Machine hearing and learning

Reminder: 15-minute talks followed by 10-minute exchanges

Coffee : 9:00 to 9:20

9:20

Classification of nonverbal human produced audio events: a pilot study

Rachel E. Bouserhal (1,4) , Philippe Chabot (1,4) , Milton Sarria-Paja (3) , Patrick Cardinal (2) , Jérémie Voix (1,4)

Ecole de technologie supérieure, Departments of (1) Mechanical Engineering and (2) Software and Information Technology Engineering (3) Universidad Santiago de Cali (4) Centre for Interdisciplinary Research in Music Media and Technology

Email: philippe.chabot.1@ens.etsmtl.ca

The accurate classification of nonverbal human produced audio events opens the door to numerous applications beyond health monitoring. Voluntary events, such as tongue clicking and teeth chattering, may lead to a novel way of silent interface command. In this pilot study, 10 nonverbal audio events are captured inside the ear canal blocked by an intra-aural device. The performance of three classifiers is investigated: Gaussian Mixture Model, Support Vector Machine and Multi-Layer Perceptron. Each classifier is trained using the mel-frequency cepstral coefficients (MFCC) and their derivatives. Fusion of the MFCCs with the auditory-inspired amplitude modulation features (AAMF) is also investigated. The highest accuracy is achieved at 75.45% using the GMM classifier with the binaural MFCC+AAMF clean training set.

9:45

An automatic mixing system for multitrack spatialization for stereo

Ajin Tom (McGill University), Joshua Reiss (Queen Mary University of London)

Email: ajin.tom@mail.mcgill.ca

One of the most important tasks in audio production is to place sound sources across the stereo field so as to reduce masking and immerse the listener in the space. This process of panning sources in a multitrack to achieve spatialization and masking minimization is a challenging optimization problem, mainly because of the complexity of auditory perception. We propose a novel panning system that makes use of multitrack sub-grouping, spectral decomposition, frequency-based spreading and an optimization algorithm to create a well spatialized mix with increased clarity while complying to the best panning practices. We also investigate objectively if this positioning strategy reduces inter-track auditory masking by using the MPEG psychoacoustic model I and various other masking metrics, extended for multitrack. Our subjective and objective tests compares the proposed work with past intelligent panning systems and human mixes. The results will be discussed.

10:10

Break the Log Jam: Per-Channel Energy Normalization Improves Mel Filterbank Feature Extraction for Speech, Music, Bioacoustics, and Sound Events

Richard Lyon (Google AI Perception)

Log-mel filterbank energy features are a staple of speech, music, and other sound processing, due to certain nice features of the logarithm. But the log has corresponding bad properties that are avoided by the recently introduced per-channel energy normalization (PCEN) approach, motivated by making the features more "auditory". With a few simple parameters, PCEN defines a space of possibilities that approach the usual log at one corner. Experiments with speech, music, bioacoustics, and sound event recognition show that the optimum is never very close to that corner.

10:35

A comparative study on filtering and classification of bird songs

Nicolas Figueiredo [1], Felipe Felix [1], Carolina Brum Medeiros [2], Marcelo Queiroz [1]

[1] Sao Paulo University, Brazil

[2] fliprl sensing

Email: carolina.medeiros@mail.mcgill.ca

We present a combination of signal processing and machine learning techniques for classification of bird song recordings. Our pipeline consists of filters to enhance the bird song signal followed by machine learning algorithms. We discuss the results of an experiment on a dataset containing recordings of bird species from South America, comparing the use of several acoustic features and three filtering techniques combined with traditional classification strategies (KNN, NB and SVM) in order to identify useful combinations for this task. This strategy produces improved classification results with respect to those reported in a previous study using the same dataset.

11:00 BREAK

11:15

Predicting verbal descriptions of algorithmic reverb presets using audio signal features

Dave Benson, Wieslaw Woszczyk, Sound Recording Area at the Schulich School of Music, McGill University

Email: david.benson@mail.mcgill.ca

In artificial reverb effects, manufacturer-chosen combinations of control parameters (i.e., presets) are typically associated with descriptive names. These names tend to contain words that describe the preset’s perceptual properties. In some cases, these words suggest general attributes of the reverb (e.g. long, short, bright, dark), while in other cases the words may indicate input signals to which the reverb is well suited (vocals, drums). In this

project, statistical learning was used to find relationships between these descriptive words and signal features of the associated reverberation, with the aim of automatically predicting words for unlabeled reverb signals. In addition to showing that predictive word models can indeed be built in many cases, the results also shed light on the acoustic signatures of certain descriptive words such as “bright”, “dark”, “drum” and “vocal”.

11:40

Melody Transcription System using Deep Neural Networks

Shayenne da Luz Moura [1], Marcelo Queiroz [1] (presented by Carolina Brum Medeiros)

[1] Sao Paulo University

Email: carolina.medeiros@mail.mcgill.ca

This work presents an implementation of a system for singing voice melody transcription based on an article published in ISMIR 2016, by François and Radenen. The main goal was to evaluate the reproducibility of articles published without source code. We coded two deep neural networks: a feed forward for F0-estimation and Bidirectional-LSTM recurrent neural network for singing/non-singing classification. Using this code it was possible to compare results training this system with different parameters and details, including some specifications not declared on the original paper. The system implemented is open source and the dataset used is freely available for research.

12:05

Lightweight Sound Source Localization, Tracking and Separation Methods for Microphone Array with Arbitrary Geometry

François Grondin, François Michaud, Introlab, Université de Sherbrooke

Email: fgrondin@mit.edu

Human-machine interaction often involves the use of microphone arrays to localize, track and separate sound sources. We present a modified sound source localization method that scans space with coarse and fine resolution grids. We also present a modified 3D Kalman method capable of simultaneously tracking the directions of multiple sound sources. Results show that the proposed methods perform at least as well as other sound source localization and tracking techniques while using up to 4 and 30 times less computing resources respectively. The proposed microphone directional model also improves separation results for a closed microphone array

12:30

Artificial Intelligence for Speech and Audio compression

Philippe Gournay, Reza LotfiDereshgi, Roch Lefebvre, Speech and Audio Research Group, Université de Sherbrooke

Email: Philippe.Gournay@USherbrooke.ca

Speech and audio compression has long been addressed from the perspective of signal processing and information theory. This has resulted in the development of a handful of key tools (for prediction, classification, transformation and quantization) and strategies (waveform matching, noise shaping) that now form the basis for virtually every standard and proprietary speech and audio coder. The ongoing remarkable development of machine learning and artificial intelligence, as well as the vast and constantly accumulating body of knowledge in

neurosciences and psychology of hearing, now suggests that a major breakthrough in the field is forthcoming. The speech and audio research group of the Université de Sherbrooke will share their vision and recent advances in developing radically new tools and strategies for speech and audio compression. The presentation will cover various theoretical and practical aspects related to a wide range of problems, from performing very specific tasks such as sound classification and signal prediction, to designing complete end-to-end artificial neural network architectures for speech and audio compression (for example based on the autoencoder model or on another generative network such as WaveNet).